INTERNATIONAL JOURNAL OF TRANSLATION
Vol.12, No.1-2, Jan-Dec, 2000

A Competition Model for the Generation of Complementation Patterns in Machine Translation

GUADALUPE AGUADO DE CEA
Universidad Politécnica de Madrid (UPM)

and

ANTONIO PAREJA-LORA
Austrian Research Institute for Artificial Intelligence (ÖFAI)

Although Fuzzy Logic has been used in Machine Translation in many ways, few applications include some kind of fuzzy syntactic module within. The purpose of the research presented in this paper was to prove that this proposal was not only possible, but also highly useful even in an interlingual approach. An interlingual machine translation method based on Lexico-Functional and Dik's Functional Grammars, applying fuzzy matching for syntactic generation, was developed and tested through the construction of a prototype for Spanish into English translation with a highly acceptable performance.

1. INTRODUCTION

Probabilities and *Fuzzy Logic* have been broadly applied to *Machine Translation* (MT) *systems* design. Models and approaches based on these disciplines are or have been used in analysis, equivalence determination and generation, the main phases of MT systems, where they have proved to be very valuable, especially to solve one of the main problems in this field: ambiguity.

Hitherto, the application of these types of mathematical and logical models (CHA93) has been increasing and improving the performance of natural language systems. All of their levels can benefit from their utilisation: the *phonological* (Morendo 1998), the *lexico-morphological* (Brown 1990), the *semantic*, as in DLT (Hutchins & Sommers 1995) when selecting the accurate word to match a given meaning, and the *syntactic*, when trying to disambiguate different possible analysis for some text fragment (López 1999).

However, little efforts have been made on applying non-deterministic, probabilistic or fuzzy logic-based strategies to the generation of target syntactic structures from the source internal representation in these MT systems.

This paper will show the results of the research carried out on how a competition fuzzy logic-based model can help a MT system to generate target syntactic structures. Especial efforts were devoted to solving one of the most complex problems in this area, that is, determining the complementation pattern of a linguistic expression in the target language, for which the solution presented in this paper may be very useful.

2. Methodology

The model presented in this paper uses two different types of verbal information. This verbal information was compiled by a large linguist team within a Spanish DGICYT funded project[1]. These two kinds of knowledge compiled were (Faber & Mairal 1999):

- Syntactic information about the different complementation patterns, that is, the specific preposition, given a verb, used for generating a target expression before its object(s) (*monolingual axes*).
  Semantic and lexical information about the correspondence between:

§ Those terms referring to a concept in the different languages studied (contrastive axes).
§ Similar concepts within a language (conceptual hierarchies).

The main aim of this project was to obtain a reusable, multilingual lexical database for Spanish, English, French and German. There was also a secondary aim: to build some applications in order to validate this lexical database and to show its utility. One of these applications was the development of a translation model based on an interlingual approach, that should be validated through a subsequent implementation. As will be shown afterwards, the competition model to be described constitutes the syntactic level of the translation model aforementioned.

2.1. The Translation Model Requirements

This translation model was developed according to the following restrictions and specifications:

- The linguistic research on which the model was based was carried out through an evolutive process. The first part of the database developed was the one dealing with Spanish and English speech act verbs. Due to this, efforts were focused initially on the design of the interlingua and the study of its behaviour and integration in the translation process from Spanish to English.

- Since the verbal category was the only grammatical category contemplated in the axes, the main aim of the validation application was to generate the accurate English verb(s) that matched the Spanish source meaning. Thus, the source expression was constrained to short sentences that included just verbs, pronouns and, of course, prepositions, in all their possible and logical combinations, depending on the complementation patterns of each verb. This allowed the inclusion of subordinate complement clauses within the source expressions considered.

- Once the terms in the target language were determined, the application should generate:

§ The exact complementation pattern, that is, the exact

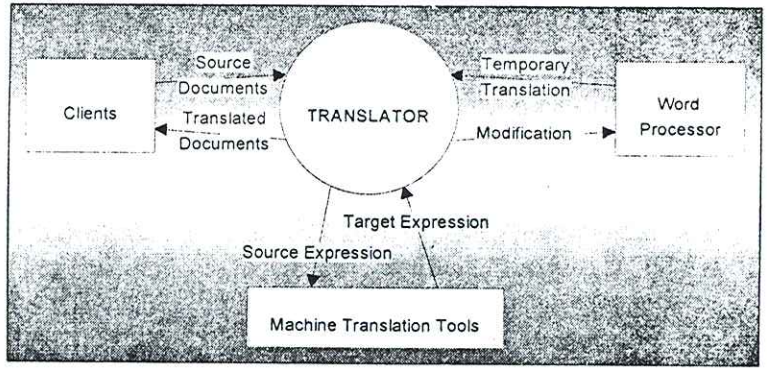
prepositional choice of the verbs involved in the target expression.

§ The precise term order in the target expression.

- Additionally, the model and also the application considered the possibility of changing the target principal verb voice from active to passive whenever its complementation pattern permitted doing so and Spanish subject was elided.

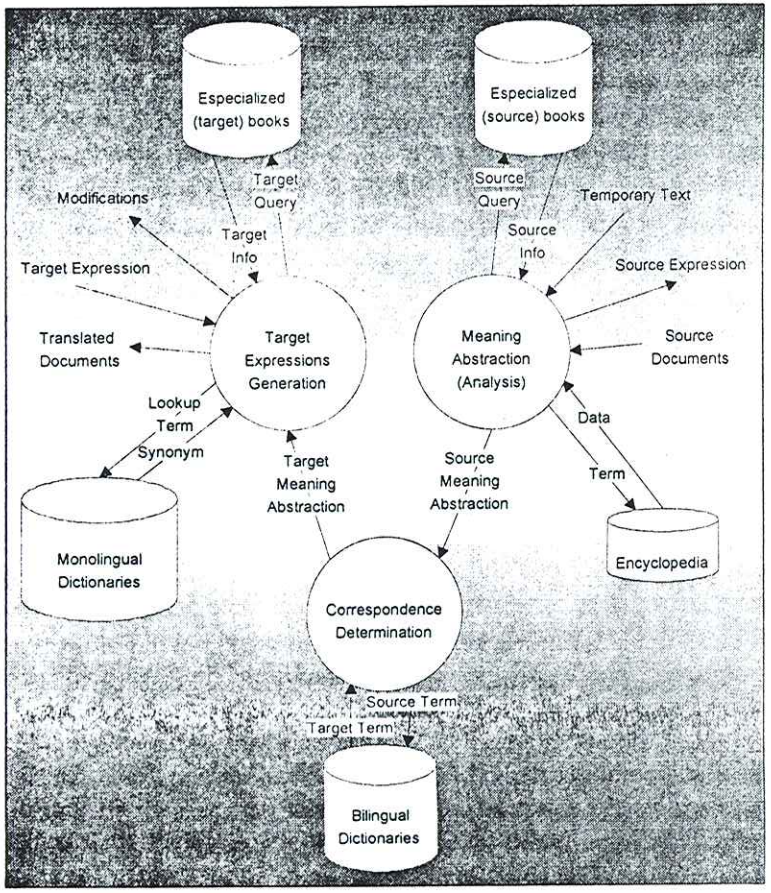
- The first prototype for the application would be developed only for two stems of the speech act verbs hierarchy: those verbs related to *singing* and those related to *ordering (commanding)*. The latter represented a particular and difficult problem to solve in MT: indirect speech expressions. In order to generalise the performance of the application, more generic verbs such as "speak", "say" or "tell" were also included in the lexicon of this first prototype.

The functional properties of the model were determined through the observation of common practices in a translation process and validated thereafter by an experienced translator. Provided that the generic context of translation processes can be depicted as shown in figure 1:



Then, the human translation process to simulate could be represented

as shown in figure 2.



2.2. The Model's Interlingua

Although presenting the interlingua itself is not the purpose of this paper, yet some lines are needed to describe the context in which our competition syntactic model operates. In an interlingual approach, the interlingua (used as a source text intermediate representation) must, as stated by Hutchins and Somers (1995), include all the necessary information for the target text generation, so that no further revision of the source text is needed. Thus, the interlingua must be

an abstract representation for both the source and the target text and, therefore, highly (if not completely) independent of the different particular language phenomena considered.

The requirements imposed on the interlingua, apart from that essential one stated above, were:

- A semantic treatment of inputs, that is, a satisfactory degree of input comprehension should be guaranteed.
- A maximal syntactic independence should be attained, in order to achieve a real semantic treatment and representation of inputs.

To ensure a high degree of independence in the interlingua, a different level of source text representation or description for each linguistic level regarded in the system was defined. Thus, the interlingua was finally designed as an aggregation of three different subinterlinguas or interlingual levels, corresponding to three major linguistic divisions:

- The semantic level, a conceptual and interlingual hierarchy;
- The syntactic level, that regards only syntactic functions, omitting further details related to structure;
- The lexico-morphological level, which establishes the correspondence between concepts in the semantic hierarchy and the different terms used in the different languages considered in the system to word each one of these concepts.

A different representation formalism was developed for each interlingual level, according to their different characteristics. Each formalism was in essence inspired in the Lexico-Functional Grammar, as it is presented in (Butler 1999) and, to some extent, in Dik's Functional Grammar (FG) (Dik 1989):

- First, the semantic level representation was substantially given in the conceptual hierarchies, built on hyponymic relations.

However, further efforts were needed to mix the already developed Spanish and English conceptual hierarchies and their contrastive axis altogether, in a unified conceptual *frame-based* (Minsky 1975) hierarchy, which bore a strong resemblance to the idea of an *ontology* as conceived in (Swartout 1997), in (Niremburg 1995) or, more generally, in (Gómez 1998). Moreover, to achieve a real semantic representation, a systematic way to state computational definitions had to be shaped, since the ones included in the axes were too close still to traditional dictionary definitions. This is whe- Dik's FG played a crucial role, with his brilliant conception of *stepwise definitions* (Dik 1986). Hence, instead of a simple conceptual hierarchy or ontology, the semantic representation in the model was conceived as a network of hierarchies or ontologies.

Second, the lexico-morphological level representation was reduced to a number of indexes for the concepts in the semantic representation, that is, a number of correspondences between terms and their (possible) meanings. This enabled the detection of lexical ambiguities for its subsequent resolution (principally at a semantic level).

Finally, an independent syntactic level representation was defined maintaining not only the pattern followed in the definition of both the Spanish and the English monolingual axes, but also including mechanisms for attaining the required level of independence from the semantic level. This level of representation and the way it is handled in generation constitute the main subject of this paper and will be presented in more detail in the next section.

2.3. The Competition Model Description

The syntactic level representation incorporated merely the syntactic function description of the lexical elements in the source or target expression, together with a brief information about ordering; the derivation of the expression meaning did not lean upon this level: it relied on the other two levels instead. The restrictions applied to the context for which the model was designed, permitted the assumption

of this hypothesis. Thus, mechanisms to obtain the syntactic interlingual representation of an expression and its correct target ordering had to be developed. These mechanisms should embody the determination of complementation patterns in order to match (analysis) or establish (generation) the syntactic functions of the expression constituents.

Being not the aim of this paper the presentation of the syntactic analysis process, but of the syntactic generation process, just a few words will be said about the former; it simply:

- Takes the lexical element list for the source expression;
- Determines the main verb;
- Tries to find its best fuzzy-matching complementation pattern for the input list, through the application of a similar competition strategy as in generation, to be explained hereafter;
- Finally, with the matched complementation pattern as a basis, it assigns its syntactic function to each element in the list.

In relation to the syntactic generation process, the steps to follow are quite the same as in analysis, but in a different order:

- The principal source verb is assumed to be the principal target verb;
- The best complementation pattern to generate the output list of lexical elements must be found, using the following competition strategy:

§ Every principal verb complementation pattern is fuzzy-matched with the list of input elements. The matching mechanism is "fuzzy" in the sense that there are multiple possible solutions and the one finally determined as the best is that which scored the highest confidence or *matching level* (N_S): the decision is not taken in terms of matching or mismatching, but in terms of matching degrees.

§ The matching mechanism is based on syntactic functions and on ordering. Whenever the syntactic function of the next input element matches the next syntactic function in the current complementation pattern, a positive increment is added to the accumulated N_S for this complementation pattern (every complementation pattern matching level is initialised with a negative value, say -1). Whenever the syntactic function of the next input element and the next syntactic function in the current complementation pattern do not match up, a negative increment is added to the accumulated N_S for this complementation pattern[2].

§ If no pattern N_S is higher than a given positive threshold (1, for example) then the principal verb is considered not to match and the system notifies the user of a syntactic mismatch.

Lastly, if a best complementation pattern match could be found, each input element is revised and properly modified, if necessary, in order to make it fulfil the characteristics imposed on it by the best complementation pattern[3].

RESULTS

To show the performance of the syntactic model explained above, the different steps taken by the prototype will be presented for one of the most complex constructions contemplated in its specifications. Let us take the case of the Spanish input sentence:

Le han ordenado hablar

which, considering that its subject (and agent) has been elided in Spanish, it should be translated into English as:

He/She has been commanded to speak

All throughout the analysis phase, the Spanish input is converted into an abstract semantic representation, based on the interlingua above mentioned. When this representation finally arrives at the syntactic generation process, it can be depicted as shown in figure 3.

Through the pointer to the conceptual frame for "command", its complementation patterns (shown in table 1, together with their

respective N_s) are obtained for their further matching[4] with the input depicted in figure 3.

SVO	2
SVO_1O_2-INF	4
SVO-That Clause	1

In the first case, SVO, a positive increment of +1 is added to Ns for each element matched; since Ns was initialised with -1, we get the value shown for this complementation pattern.

In the second one, the matching of the second object entails the adding of +2 to the accumulated value of *Ns*. The fourth element does have not only an object syntactic function, but it is also an infinitive (see figure 3).

As can be observed, the third case is the opposite to the last mentioned: *SVO-That Clause* is penalised with a negative (-2) adding, since the input does not match with the requirement that the object should be a clause introduced by "that".

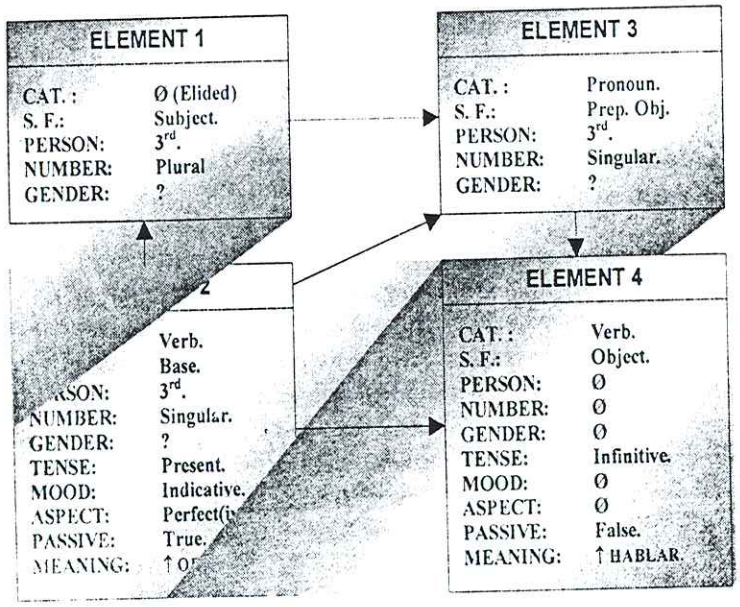


Figure 3: Interlingual Representation for *"Le han ordenado hablar"*

Since the second complementation pattern renders the highest score (*Ns* = 4) and is higher than the pre-determined threshold, hence it is the one that "survives" (as in Darwin's theory) and wins the competition, which is the result expected. The order of the elements is corrected in the revision of the input in order to fulfil the characteristics imposed on it by this complementation pattern, taking into account that a passive voice must be generated. The element list obtained in this way is given to the morpho-lexical generation process, which returns the desired translation.

3. Conclusion

The multiple syntactic differences across languages make it difficult to determine a real interlingual representation for machine translation. Hence, a maximum level of syntactic independence must be introduced in the input semantic analysis and representation of these systems in order to guarantee a good final semantic translation. In this paper, a method combining disparate areas such as (Lexico-Functional) Grammar, Fuzzy Logic and Ontological Engineering has been proposed for solving this particular problem.

The performance of the method, shown in the results section, gives support to the hypothesis that a sufficient level of independence in syntactic generation can be achieved in machine translation interlingual approaches, at least in the domain of speech act verbs, for which the prototype was developed. This simply requires additional mechanisms to the ones commonly used. In the model presented in this paper, the syntactic differences between the possible language couples are overcome by the introduction, in the system, of knowledge related to complementation patterns and fuzzy matching. However, further research should be done to test this method in other semantic fields like movements, feelings, etc., where the linguistic studies already developed have not been examined in the light of this model. The restrictions and requirements derived from the inclusion of other languages in the interlingua syntactic level must be also determined, so that the method could attain a satisfactory degree of generality.

References

Brown, P. J., Cocke, S . et al. A Statistical Approach to Machine Translation" In: *Computational Linguistics 16*. 1990.

Butler, C., Mairal, R. et al. *Nuevas perspectivas en Gramática Funcional*. Editorial Ariel. Barcelona, 1999.

Dik, Simon C. Functional Grammar and its potential computer applications. In: *Two papers on the computational applications of Functional Grammars*. Amsterdam, 1986. Amsterdam: *Working Papers in Functional Grammar 18*.

Dik, Simon C. *The theory of Functional Grammar*. Foris Publications. Dordrecht, 1989.

Faber, P. & Mairal Usón, R. *Constructing a Lexicon of English Ver*bs. Berlin:MoutondeGruyter, 1999

Gómez-Pérez, A. *Nociones de Ingeniería Ontológica* (*What is Ontological Engineering*?). Publication Service, Facultad de Informática, UPM. Madrid, 1998.

Hutchins W.J. y Sommers H.L. *Introducción a la Traducción Automática*. Editorial Visor. Madrid, 1995.

López-Soto, M.T. Aplicación Estadística a la Traducción Automática. In: *Philologia Hispalensis: Procesamiento del Lenguaje Natural*. Editorial Kronos. Sevilla, 1997.

Minsky, M. A framework for representing knowledge. In: *The Psychology of Computer Vision*. McGraw-Hill. New York, 1975.

Mori, Renato De & Brugnara, Fabio. HMM Methods in Speech Recognition. In: *Survey of the State of the Art in Human Language Technology*. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html

Moreno-Sandoval, A. *Lingüística Computacional*. Editorial Síntesis. Madrid, 1998.

Niremburg, S., Raskin, V., Onyshkevych, B. *Apologiae Ontologiae*. NMSU CRL MCCS-95-281.

Ruiz-Antón. J.C. "La Gramática Funcional de Dik como módulo de generación sintáctica" in *Philologia Hispalensis: Procesamiento del Lenguaje Natural*. Editorial Kronos. Sevilla, 1997.

Swartout, B., Patil, R., Knight, K., Russ, T. "Toward Distributed Use of Large-Scale Ontologies" in *Ontological Engineering*. AAAI-97 Spring Symposium Series, 1997.

1 DGICYT Project PB94/0437, coordinated by Dr. R. Mairal-Usón.

2 Principal verbs are excluded from the matching process: they are assumed to match (+1 adding).

3 This approach can be compared to a *crisp* (not fuzzy) and PROLOG based one in [RUI99].

4 CAT. and S.F. stand for grammatical category and syntactic function, respectively; a ø value for GENDER means that it could not be determined during the lexical analysis. Whenever something like ORDENAR appears, it must be understood as a pointer to the verbal concept corresponding to "ORDENAR" in the conceptual hierarchy, which coincides with the one corresponding to its translation (in this case, "COMMAND").

GUADALUPE AGUADO DE CEA
Dept. Lingüística Aplicada a la Ciencia
y a la Tecnología (DLACT)
Universidad Politécnica de Madrid (UPM)
e-mail: lupe@fi.upm.es

ANTONIO PAREJA-LORA
Austrian Research Institute for
Artificial Intelligence (ÖFAI)
e-mail: apareja@ai.univie.ac.at